BAB II

LANDASAN TEORI

2.1 Text Preprocessing

Text Preprocessing adalah suatu proses pengubahan bentuk data yang belum terstruktur atau tidak terstruktur menjadi data yang terstruktur. Tujuannya adalah untuk memperkecil dimensi data sehingga proses komputasi lebih menjadi efisien dan diharapkan lebih presisi. Sebuah teks yang ada harus dipisahkan, hal ini dapat dilakukan dalam beberapa tingkatan yang berbeda. Suatu dokumen dapat dipecah menjadi bab, sub-bab, paragraf, kalimat dan pada akhirnya menjadi potongan kata/token. Selain itu pada tahapan ini keberadaan digit angka, huruf kapital, atau karakter-karakter yang lainnya dihilangkan dan diubah. *Preprocessing* terdiri dari beberapa tahapan. Adapun tahapan *preprocessing* berdasarkan, yaitu: *case folding, filtering, dan stemming* (Pramudita, et al., 2018).

1. *Case Folding* merupakan tahapan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil (lowercase). Huruf yang dilakukan perubahan mulai ‘a’ sampai dengan ‘z’.
2. *Tokenizing* merupakan proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter seperti tanda titik (.), koma (,) dan spasi serta karakter yang ada pada kata tersebut.
3. *Stemming* adalah suatu proses untuk mereduksi kata ke bentuk dasarnya. Tahap stemming merupakan tahap mencari akar (root) kata dari tiap kata hasil filtering (Agus, et al., 2015).

2.2 Naïve Bayes Classifier (NBC)

Klasifikasi merupakan sebuah proses penentuan model maupun fungsi yang membedakan konsep atau kelas data dengan tujuan memperkirakan kelas yang tidak tersedia pada objek tersebut dan terdapat 2 proses yang dilakukan pada saat klasifikasi yaitu:

1. *Training*

   Proses ini dilakukan training set yang sudah diketahui label-labelnya untuk membangun model.

2. *Testing*

   Proses ini intuk mengetahui keakuratan model yang dibangun pada proses training, umumnya digunakan data yang disebut test set untuk memprediksi label.

Metode NBC terdiri dari dua tahap dalam proses klasifikasi teks, tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sample dokumen berupa pemilihan *vocabulary* yaitu kata yang dimungkinkan muncul dalam koleksi dokumen sampel yang menjadi representasi dokumen. Langkah selanjutnya adalah menentukan probabilitas bagi tiap kategori berdasarkan sampel dokumen (Wijaya dan Santoso, 2016). Teorema Bayes merupakan teorema yang mengacu konsep probabilitas bersyarat. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi. Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_2, a_3,..., a_n", dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Secara umum teorema *Bayes* dapat dinotasikan pada persamaan berikut:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \ldots(6.1)$$

Sedangkan V adalah himpunan kategori berita. Pada saat klasifikasi algoritma mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun persamaan V_{MAP} adalah sebagai berikut:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j) \quad \ldots(6.2)$$

Nilai $P(v_j)$ dihitung pada saat data *training*, didapat dengan rumus sebagai berikut:

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad \ldots(6.3)$$

Dimana $|doc\ j|$ merupakan jumlah dokumen (artikel berita) yang memiliki kategori *j* dalam *training*. Sedangkan |training| merupakan jumlah dokumen (artikel berita) dalam contoh yang digunakan untuk training. Untuk probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$ dihitung pada saat training.

$$P(a_i|v_j) = \frac{n_i+1}{doc\ j} \quad \ldots(6.4)$$

Dimana n_i adalah jumlah kemunculan kata a_i dalam dokumen yang berkategori v_j, sedangkan n adalah banyaknya seluruh kata dalam dokumen dengan kategori v_j dan $doc\ j$ merupakan jumlah dokumen (artikel berita) yang memiliki kategori j dalam *training*. Pada penelitian ini menggunakan algoritma *Bernoulli Naïve Bayes*. Pada *Bernoulli Naïve Bayes*, pembobotan dilakukan dengan menggunakan binary (0 dan 1) dalam pembobotan tiap *term*, hal ini berbeda dengan perhitungan term frekuensi yang melakukan pembobotan pada setiap *term*.

Adapun beberapa bentuk representasi dari metode *Naïve Bayes* selain *Bernoulli* antara lain:

1. *Gaussian Naïve Bayes*

   Ketika berhadapan dengan data yang berkelanjutan, asumsi tipikal adalah bahwa nilai berkelanjutan yang terkait dengan setiap kelas terdistribusikan sesuai dengan distribusi *Gaussian* (Gayathi & Sumathi, 2016).

2. *Multinomial Naïve Bayes*

   Multinomial Naïve bayes mengasumsikan indepedensi diantara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan kata dan konteks informasi dalam kalimat atau dokumen secara umum. Selain itu metode ini memperhitungkan jumlah kemunculan kata dalam dokumen (Destuardi and Sumpeno, 2009).

2.3 N-Gram

Bahasa tidak terbentuk dari kata-kata individu, tetapi terdiri dari urutan kata individu dan frase 2, 3 atau lebih kata yang lebih dikenal *n-gram* dengan masing-masing kata tersebut mengandung informasi tersendiri (Pujadayanti, et al., 2018). Model *n-gram* adalah salah satu teknik statistik yang memodelkan urutan kata kedalam nilai probabilitasnya. Teknik *n-gram* didasarkan pada pemisahan teks menjadi string dengan panjang n mulai dari posisi tertentu dalam suatu teks. Posisi *n-gram* berikutnya dihitung dari posisi yang sebenarnya bergeser sesuai dengan offset yang diberikan. Nilai offset bergantung pada pembagian yang digunakan dalam *n-gram*. Pembagian *n-gram* dapat bervariasi tergantung dari pendekatan dalam membagi teks menjadi bentuk *n-gram*. *N-gram* untuk setiap string dihitung dan kemudian dibandingkan satu per satu. *N-gram* dapat berupa unigram (n=1), bigram (n=2), trigram (n=3), dan seterusnya (Lisangan, 2013).

Penelitian ini menerapkan n-gram dengan pemecahan kata pada kalimat ulasan meliputi bigram yang merupakan pemecahan *n*-kata pada kalimat ulasan dengan *n*=2. Berikut ilustrasi penerapan *n*-gram dengan *bigram* pada kalimat "aku suka banget produk cleanser ini." (Prasanti, et al., 2017): 'aku suka', 'suka banget', 'banget produk', 'produk cleanser', 'cleanser ini'.